A Machine Learning Univariate Time Series Model for Forecasting COVID-19 Confirmed Cases : A Pilot Study in Botswana

Ofaletse Mphale[1], Ezekiel U Okike[2], and Neo Rafifing[3]

[1,2,3] Department of Computer science, University of Botswana, Plot 4775 Notwane Rd, Gaborone, Botswana

Abstract
The recent outbreak of corona virus (COVID-19) infectious disease had made its forecasting critical cornerstones in most scientific studies. This study adopts a machine learning based time series model - Auto Regressive Integrated Moving Average (ARIMA) model to forecast COVID-19 confirmed cases in Botswana in 60 days period. Findings of the study confirmed COVID-19 cases steadily rising with trend of random fluctuations and non-constant variance. This trend could be effectively described using an additive model in Seasonal Trend Decomposition procedure using Loess (STL). In selecting the best fit ARIMA model a Grid Search Algorithm (GSA) was used. The Akaike Information Criterion (AIC) metrics was used to derive scores of the different fit ARIMA models. In this study ARIMA model (5, 1, 1) with corresponding AIC score of 3885.091 was nominated. This model depicted the least value of the AIC measure. The forecasts results obtained from the study proved that ARIMA model could efficiently provide reliable estimates of forecasts that could be used to guide on understanding of future spread of COVID-19 confirmed cases. Findings of the study were useful in raising social awareness to disease monitoring institutions and government regulatory bodies where it could be used to support strategic health decisions and initiate policy improvement procedures for better management of the COVID-19 pandemic..

Keywords:
COVID-19, Corona virus, ARIMA, Box-Jenkins, Time series, Machine learning, ACF, PACF, AIC

1. Introduction

Coronaviruses are infectious diseases that are closely related to common cold, Middle East Respiratory Syndrome coronavirus (MERs) and Severe Acute Respiratory Syndrome coronavirus (SARs). These diseases are once known to diffuse from animals to human beings. For example; SARs, was known to transfuse from civet cats to humans while MERs was transmitted to humans from a type of camel (Singla, Mishra, & Joshi, 2020). The corona virus was officially termed as "COVID-19" by the World Health Organisation (WHO) and its first incident was registered in Wuhan city in China on December 2019. Since then, the virus had spread rapidly reaching different segments of the world (BBC News, 2020) (Our World in Data, 2020) (Vara, 2020).

Fatalities from COVID-19 had been presented in amplifying figures globally. In recent findings, it had been shown that COVID-19 fatalities had surpassed 1.9 million, with confirmed cases exceeding of 88.5 million worldwide (Worldometer, 2021) (WHO, 2021). In state-of-art, studies had shown that there had been limited attempts in clinical trials conducted to evaluate potential COVID-19 treatments (IWK Health Center, 2021) (Hodgson, Mansatta, & Mallet, G. et al, 2020). However, some crucial recovery measures had been outlined such as; self-isolation, drinking of plenty of water, consumption of paracentamol and adequate rests (WHO, 2021)

COVID-19 disease can affect individuals in different age spectrums. It is transfused through human to human contact. Its symptoms are characterised by diseases like flu, fever, fatigue and respiratory complication. Elderly people with other chronic diseases like diabetes and high blood pressure are the most vulnerable to the undesirable effects of the disease. Some safe guard measures such as frequent hand wash, wearing face mask and social distancing had been suggested in some studies to reduce contamination with the disease (Ceylan, 2020) The consequences of infectious diseases do not only detriment human health, but it is also an economic burden. With most countries employing procedures to control the virus such as lock down, mobility restrictions, quarantine and more, however the accurate prediction of infectious diseases remains a global challenge. Furthermore, forms of occurrences of infectious diseases are often unknown (WHO, 2021).

Time series models are popular machine learning techniques applied in different scientific grounds to discover trends and relationships in series data. This study adopts ARIMA model to forecast corona virus confirmed cases in Botswana over two months period. ARIMA model follows a Box-Jenkin approach for time series forecasting. These findings are foreseen to raise social awareness to disease monitoring institutions and the government regulation bodies where it could be utilised to support

Manuscript received December 5, 2024
Manuscript revised December 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.12.13

strategic health decisions and enhance policy improvement procedures, for better management of the COVID-19 disease..

The rest of the paper is organised as follows; Section 2 presents the literature review of the subject being studied. That is different theoretical and empirical scholars' perceptions on application of ARIMA models in modelling and forecasting of infectious diseases. In Section 3, methodology framework to be followed by the study analysis process is presented. Section 4 presents the Results and Discussion of the study. Finally conclusions and future works are discussed in section 5.

1 Literature Review

With the recent advancements of forecasting methods like machine learning, artificial intelligence and mathematical models, scholars had come to appreciate them and had integrated them in different studies to tackle real world tasks. Predictive analytics learn from historical data and utilises machine learning approaches to derive future conclusions. The application of machine learning algorithms in technical grounds like engineering, computer science, medicine, statistics etc. had made it possible to recognize infectious diseases patterns, accelerate diagnosis in order to forecast their future directions.

Infectious diseases are caused by pathogenic microorganisms such as bacteria, viruses, parasites or fungi which are diffused between individuals or an animal (Steptoe & Poole, 2015). Zoonotic diseases are groups of infectious diseases that affect animals, but can cause diseases when transmitted to humans (World Health Organization, 2021).Studies had shown that to date, various models and tools had been developed to predict and monitor outbreaks of infectious diseases. In a study (Chaurasia & Pal, 2020) compared different forecasting methods such as Holt linear trend method, naive method, single exponential smoothing, simple average, Holt-Winters method , moving average and ARIMA using root mean square error score. In their findings it was deduced that the naïve model outperformed all other models. However, based on the ARIMA model, the grid search method yielded the best fit model for the series data. Furthermore it was concluded that the number of COVID-19 deaths will surpass 600 000 in January 2021.

In the kingdom of Saudi Arabia (Abuhasel, Khadr, & Alquraish, 2020) applied classical SIR model to predict the highest number of COVID-19 cases that could be recognised to flatten the curve. Similarly ARIMA model was used to predict the prevalence cases of COVID-19. In their findings it was it was deduced that the SIR model affirmed that the containment technique used by Saudi Arabia to curb the spread of the disease was efficient. By validating the performance of the applied models, ARIMA proved to be a good forecasting method from current data. In another study (Chae, Kwon, & Lee, 2018) optimised deep learning parameters to predict outbreak of infectious diseases such as chicken pox. The study made use of social media big data. Predictive models such as ARIMA, deep neural network (DNN) and long-short term-memory (LSTM) were investigated. Based on the results, it was established that DNN and LSTM models perform better than ARIMA. However, LSTM model produced more accurate predictions compared to DNN model in particularly modelling disease outbreak.

In a recent study, mathematical models such as Logistic model, Bertalanffy model and Gompertz model had been examined using SARs epidemic trends data. Findings of the study had shown that the three models performed differently with different parameters in different regions (Jia, Li, Jiang, Guo, & zhao, 2020). When evaluating the forecasting models; ARIMA, LSTM, back-propagation artificial neural network (BP-ANN) and seasonal trend decomposition using Loess + ARIMA on malaria data from Yunnan Province (Wang, Wang, Wang, & Lui et. al, 2019) deduced that the four models performed better stacked with gradient-boosting regression trees. In the study findings it was concluded that assemble algorithms could improve prediction of the infectious diseases prediction models.In a study (Wang, Xu, Zhang, & Yang et. al, 2019) modelled seasonal patterns of hand, foot, and mouth disease (HFMD) in children in mainland China. The study adopted LSTM time series model to make predictions. Moreover, the prediction performance of the LSTM was compared against ARIMA and auto-regressive neural network models. Based on findings it was deduced that LSTM had the best fitting and better forecast performance compared to the other models. Results from HFMD trends showed rising trends in summer signifying high-risk season. Furthermore, it was concluded the LSTM method could be relevant in predicting outbreak of malaria disease incidents.

2 Methodology

2.1 Dataset Description

The data set used in the study was acquired from John Hopkins University Web portal and other relevant sources (Worldometer, 2021) (Johns Hopkins University of Medicine, 2021) (AccuWeather, 2021). These are data repositories which are freely available for public use for academic and non-academic purposes. Therefore, the acquired data set consisted of COVID-19 global registered cases (Confirmed, Recovered, Deaths) from 31st November

2019 to 12th January 2021. Since the study was only interested in investigating COVID-19 confirmed cases in Botswana, some observations and attributes were pruned from the final dataset. The IBM Statistical Package for the Social Sciences and R studio software were used to analyse the set data.

In data pre-processing steps some processes were applied, for example; feature label transformations, replacing missing values (confirmed cases which were left blank were replaced with a '0' value), removal of outliers (characters and some negative values representing confirmed cases were replaced with an average value derived from the dataset), transformation of the series attributes to appropriate data types (the Date attribute in dataset was transformed to Date data type and the confirmed cases attribute data type was set to numeric attributes accordingly). Lastly the series was transformed to proper time series data frame for further analysis. Therefore the final dataset comprised of only confirmed cases in Botswana registered from month of 4th April 2020 to month of 12th January 2021. The proposed methodology framework adopted for the study analysis process is illustrated in Figure 1 as shown.

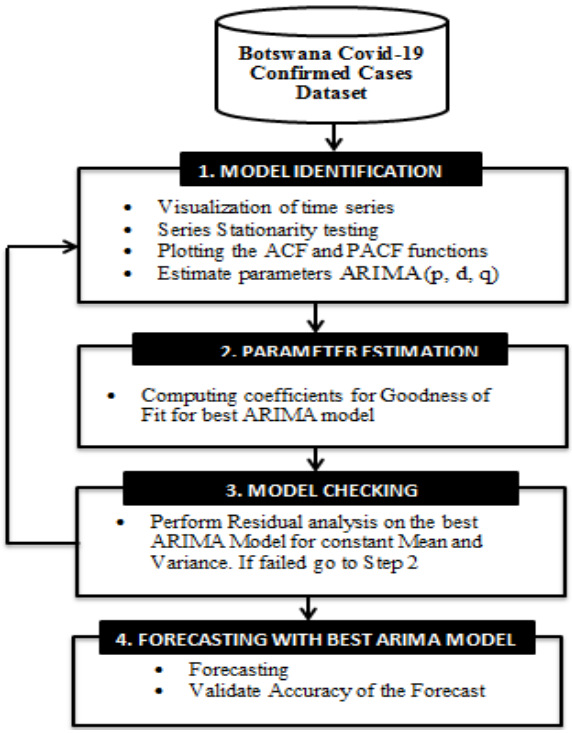

Fig. 1. Proposed methodology framework for the analysis process of the study

2.2 Time Series Analysis

Time series is simply expressed as a set of data points ordered in time (Fanoodi, Malmir, & Jahantigh, 2019). It is assumed to compose of random walk (non-stationary series) and white noise (zero mean stationary series). In mathematical terms time series is defined as shown;

$$\mathbf{y_t} = f(t) \quad (1)$$

Where, $\mathbf{y_t}$ presents the value of the variable under study at time **t**.

If the population is the variable being studies at various time period $t_1, t_2, t_3, \ldots, t_n$.
Then the time series is briefly elaborated as shown;

$$t: t_1, t_2, t_3, \ldots, t_n$$
$$\mathbf{y_t}: y_{t1}, y_{t2}, y_{t3}, \ldots, y_{tn}$$

or,

$$t: t_1, t_2, t_3, \ldots, t_n$$
$$\mathbf{y_t}: y_1, y_2, y_3, \ldots, y_n \quad (2)$$

Time series forecasting approaches could be categorised in to two broad classes. These are univariate time series forecasting and multivariate series forecasting. In univariate series forecasting, predictions of future data points ultimately depends on previous values in the series while in multivariate time series analysis other predictors (exogenous variables) other than the series values are taken in to account in forecasting. In this study, based on the nature of the dataset that was examined, a univariate time series analysis approach was deemed appropriate to model forecasts.

The subsequent step was to establish whether the series decomposition was additive or multiplicative. In an additive time series, seasonality and residuals are independent of the trend whereas in multiplicative time series is vice versa. The mathematical description of an additive time series is given as shown in Equation (3);

$$O_t = T_t + S_t + R_t \quad (3)$$

Where,

$\boldsymbol{O_t}$ – represents the output
$\boldsymbol{T_t}$ – represents the trend
$\boldsymbol{S_t}$ – represents the seasonality
$\boldsymbol{R_t}$ – represents the output

Similarly, in multiplicative time series, the mathematically description could be given as shown;

$$O_t = T_t * S_t * R_t \quad (4)$$

This study adopted an additive time series decomposition analysis. Then ARIMA model was then developed to generate forecast of the daily confirmed cases in Botswana.

2.3 ARIMA Algorithm

ARIMA is a statistical machine learning based algorithm used in time series forecasting. It was discovered by statisticians; George Box and Gwiltyn Jenkins. It is also known as Box-Jenkins model. ARIMA model extends AR (Auto Regressive) and MA (Moving Average) models by integrating with order of differencing steps. It relies on the known historical data to establish future forecast values (Fattah, Ezzine, & Aman, 2018).

In order to successfully apply the ARIMA model, a non-stationary time series must be transformed from random walk to white noise. Random walk series trends produce unreliable forecasts. Stationary series has a constant mean, variance and its autocorrelation structures are not affected by fluctuations over time. To test for stationarity, an Augmented Duckey Fuller (ADF) test could be used. ADF test examines the null hypothesis for the presence of a unit root in series. It is described shown in Equation (5);

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} .. + \phi_p \Delta Y_{t-p} + e_t \quad (5)$$

Where,

- y(t-p) = lag ***p*** of time series
- delta Y(t-p) = ***p*** difference of the series at time (t-p)

The existence of unit root in series makes it non-stationary. If the unit root exists in series, then it is recognised at value of alpha (α) = 1. The criterion used to test the null hypothesis for presence of the unit root is given as shown:

Given alpha is = 1;

H0: The series has a unit root.
H1: The series does not have a unit root. The series is stationary.

To assess the presence of unit a root in series, the p-value obtained should be less than the significance level of 0.05. In that way the null hypothesis is rejected otherwise it is affirmed that the time series is non stationary.

Mathematically, ARIMA model is defined as shown; **ARIMA (*p, d, q*)**

Where,

p – Represents the order of AR polynomial indicating of the autoregressive model lags
d – Represents the order of the differencing
q – Represents the MA polynomial order of the moving-average process

If the series is already stationary, then ARIMA model can be presented as an ARMA (***p, q***) with a differencing sequence of ***d*** times, where ***p, d, q*** ≥ 0. It is simplified as ARMA model as shown;

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j}, \quad (6)$$

Where $\phi_1, \ldots, \phi_p$ are the AR parameters to be estimated, $a_1, \ldots, a_t$ are the MA parameters to be estimated and are a series residuals that follows a normal distribution. The equation 3 could be further simplified by applying the Box-Jenkins backshift operator as shown;

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) Y_t = \left(1 - \sum_{j=1}^{q} \theta_j B^j\right) a_t,$$

Then equation 7 could be further condensed to Equation (8) as shown;

$$\phi_p(B) Y_t = \theta_q(B) a_t, \quad (8)$$

Where,

$$\phi_p(B) = (1 - \sum_{i=1}^{p} \phi_i B^i) \text{ and } \theta_q(B) = (1 - \sum_{j=1}^{q} \theta_j B^j).$$

If the series is non stationary, then then ARIMA model can further extended by integrating with the order of differencing steps as illustrated as shown;

$$\begin{aligned} W_t &= Y_t - Y_{t-1} = (1 - B) Y_t \\ W_t - W_{t-1} &= Y_t - 2Y_{t-1} + Y_{t-2} \\ &= (1 - B)^2 Y_t \\ &\vdots \\ W_t - \sum_{k=1}^{d} W_{t-k} &= (1 - B)^d Y_t, \end{aligned} \quad (9)$$

Where ***d*** is the order of **differencing Steps**. Replacing the ***Yt*** in the ARMA model with the differences defined in Equation (9) then the formal definitions **ARIMA (*p*, *d*, *q*)** model could be simplified as shown in Equation (10);

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)a_t. \quad (10)$$

Testing for Goodness of Fit

Testing for goodness of fit validates whether ARIMA models are an appropriate fit to the data. In this study an AIC metric was used to evaluate the model fit scores of the ARIMA model. This was executed using GSA developed in python language. The lowest value of the AIC denoted a good model fit. AIC could describe in the following Equation (11);

$$\textbf{AIC} = -2(\text{log-likelihood}) + 2k \quad (11)$$

Where;

- ***k*** is the number of model parameters (the number of variables in the model plus the intercept).
- Log-likelihood is a measure of model fit. The higher the number, the better the fit. This is usually obtained from statistical output

For ARIMA models the AIC equation could be rearranged as shown;

$$\text{AICc} = \text{AIC} + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}, \quad (12)$$

Residual analysis for the fitted ARIMA model

The Ljung-Box test was used to test lack of good fit. This examined behaviour of residual after fitting an ARIMA model of the observable series. If it was learnt that the auto correlations of the residues are small, then the model has insignificant lack of good fit. The Ljung-Box test is given as shown;

$$Q(m) = N(N+2)\sum_{h=1}^{m}\frac{\widehat{\rho}_h^2}{N-h}. \quad (13)$$

Where,

- $\widehat{\rho}_h$ is the estimated autocorrelation of the series at lag k,
- *m* is the number of lags being tested.

Alternatively, statistical hypothesis testing of the Ljung-Box test is given as shown;

H0: The model does not exhibit lack of fit.
Ha: The model exhibits lack of fit.

For significance level α, the critical region for rejection of the hypothesis of randomness in Ljung-Box test is given as shown;

$$Q > \chi^2_{1-\alpha,h} \quad (14)$$

Where, $\chi^2_{1-\alpha,h}$ is the 1-α quantile of the chi-squared distributio degrees of freedom.

Evaluation of accuracy of estimated forecasts

There are various error metrics that could be utilised to assess the accuracy of the forecasts in ARIMA models. In this study Ljung-Box statistical test error metrics was used. Thus; a successions of Ljung-Box tests were run on forecasts residuals with different lag values. The sizes of the residuals were observed. The statistical hypothesis tests of the Ljung-Box were used to stem various conclusions.

3 Results and Discussions

The major objective of the study was to model forecasts of COVID-19 confirmed cases in Botswana using ARIMA Box-Jenkin model. Therefore in order to develop the most suitable ARIMA Box-Jenkin model the study followed the four main stages. Thus; model identification, model parameter estimation, diagnostic checking and forecasting with best ARIMA model. These are also given in section 3 – Methodology.

3.1 Model Identification

In this phase, the series for COVID-19 confirmed cases was plotted and its various components such as seasonality, trends and noise were analysed. Plotting the series is an essential way to gain preliminary understanding of the series structure and an initial step to determine the most suitable forecasting model for the data (Abuhasel, Khadr, & Alquraish, 2020). Figure 2 presents the graphical illustration of the COVID-19 Confirmed Cases in Botswana as shown.

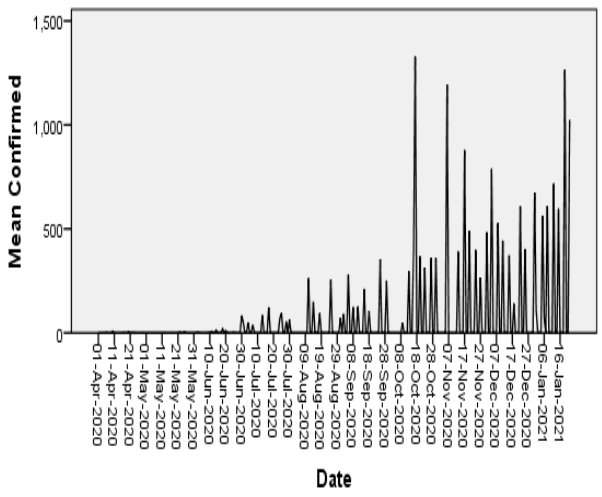

Fig. 2. COVID-19 Confirmed Cases in Botswana over Time

Fig. 2 shows that from April 2020 to January 2021, COVID-19 confirmed cases in Botswana have been rising steadily with a horizontal trend and some periodic spikes. The series also depicts daily cycles of fluctuations that revert around zero mean. Furthermore, the series depicts components of a weak stationarity which could be effectively decomposed using additive model. In period of the month of April 2020 up to month of early October 2020 results showed that there were minimum figures of COVID-19 confirmed cases registered in Botswana. These results could be related to the effectiveness of government policies and precaution measures which were imposed e.g. national lock down, quarantine, restriction of movements, compulsory wearing of face masks in public areas and more.

In late October 2020, results showed that the series had registered its first sharpest spike. This was during the early period of state of emergency extension in the country. Hence, this could indicate that most precaution measures were still under consideration or were just implemented. However the confirmed cases figures degraded towards the month of December 2020. Then towards the month of January 2021 confirmed cases rise in steadily trend. This could also suggest that the government precaution measures which were imposed in that period such as curfew, regulation of liquor stores trading hours, social distancing, prohibition of public gatherings and more, had a slight effect towards controlling COVID-19 daily infections.

In order to verify that ARIMA model was indeed an appropriate model for the data. The Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots were fitted with the data. The behavioural patterns of the correlograms were investigated in respective ACF and PACF plots. Figure 3 presents the graphical illustration of the results of confirmed cases fitted in ACF and PACF plots as shown;

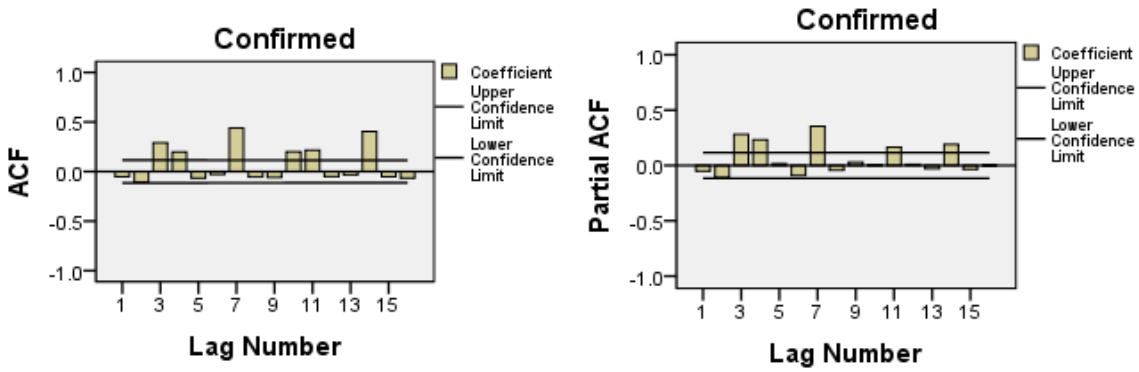

Fig. 3. ACF and PACF correlograms plots of confirmed cases in Botswana

Fig 3 shows that the both ACF and PACF correlograms follow the same pattern. Furthermore both plots are show statistical significance at lag 3, 4, 7 and 11 respectively. This also indicates a series that could be modelled appropriately using an ARIMA processes.

In order to gain further understanding of the series trend and its seasonal properties, the next step was to decompose the series data to separate trend component and random component. In this study the confirmed cases data set was decomposed using an additive model following STL procedure. Figure 4 depicts the graphical illustration of series decomposition in STL as shown;

Fig. 4. The STL series Decomposition graph of COVID-19 Confirmed cases in Botswana over Time

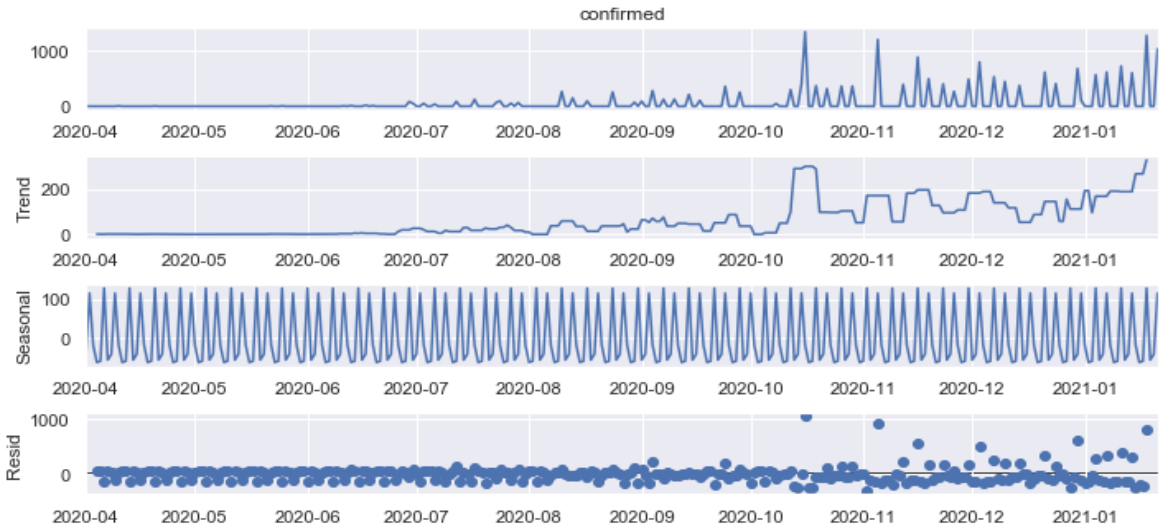

Fig. 4. The STL series Decomposition graph of COVID-19 Confirmed cases in Botswana over Time

Fig.4 shows that the trend COVID-19 confirmed cases in Botswana are increasing constantly with a fluctuating upward trend over time. The series also show daily cycles of patterns of relative size over time. In the residual plot, the series show that residues had been constant in the early months of the series. That is from the beginning of April 2020 until September 2020. However, after October 2020 the residuals had shown variations of random noise. This also described characteristics of non stationarity series. In a stationary time series, the residual distribution are assumed to compose of variance and covariance that revert around zero mean (Koopmans, 1995).However, in order to confirm that the series was stationary or not, an ADF statistical test method was used. This examined the null hypothesis for the presence of unit root at alpha (α) = 1.

Testing for stationarity

Based on the ADF statistical test results, the p-value measure obtained was 0.330804. This was greater than the threshold significant level of 0.05. In that way the null hypothesis was not rejected. Therefore, the study accepted the alternative hypothesis. The synopsis of the ADF stationarity test results was given in Figure 5 as shown;

```
Results of Dickey-Fuller Test:
Test Statistics                -1.902675
p-value                         0.330804
#Lags Used                     16.000000
Number of Observations Used    334.000000
Critical value(1%)             -3.450081
Critical value(5%)             -2.870233
Critical value(10%)            -2.571401
dtype: float64
```

Fig. 5. The ADF stationarity test results

Fig.5 shows that at maximum lags of 16, the critical value was -1.902675. This was still greater than the critical value of -2.870233 at 5% confidence interval. Therefore the null hypothesis was rejected. Since ARIMA modeling requires non-stationary series to be converted to white noise, the next step was to generate a stationary series. In this step different combination of ARIMA models parameters (p, d, q) were fitted with the data model parameter estimation phase was conducted.

3.2 Model Parameter Estimation

In this phase, the series variance was initially normalised by applying log transformation procedure. Then a GSA was run recursively on different combinations of ARIMA model parameters. In order to ensure that the ARIMA models would not over fit, ARIMA parameters were limited to a recursive range between 0 and 6. The

models were optimised for goodness of fit using the maximum likelihood estimation procedure. In this study the lowest score of AIC metric was used to determine the best fit ARIMA model. Figure 6 presents an overview of the GSA results from fitting different ARIMA models and their corresponding AIC scores as shown.

ARIMA Models	AIC Scores
ARIMA(2,1,2) with drift	3887.532
ARIMA(0,1,0) with drift	4151.456
ARIMA(1,1,0) with drift	4083.781
ARIMA(0,1,1) with drift	3916.383
ARIMA(0,1,0)	4149.502
ARIMA(1,1,2) with drift	3902.917
ARIMA(2,1,1) with drift	3887.491
ARIMA(1,1,1) with drift	3909.989
ARIMA(2,1,0) with drift	3971.975
ARIMA(3,1,1) with drift	3886.553
ARIMA(3,1,0) with drift	3925.584
ARIMA(4,1,1) with drift	3884.729
ARIMA(4,1,0) with drift	3925.421
ARIMA(5,1,1) with drift	3885.091
ARIMA(4,1,2) with drift	3886.184
ARIMA(3,1,2) with drift	3887.049
ARIMA(5,1,0) with drift	3928.319
ARIMA(5,1,2) with drift	3885.156
ARIMA(4,1,1)	3888.293

Fig. 6. The GSA result of fitting ARIMA models and corresponding AIC Scores

Fig.6 shows that the best ARIMA model fit is at ARIMA (5, 1, 1). This model showed the AIC measure of 3885.091 which was the lowest AIC score in complete GSA runs for the goodness of fit. The ARIMA model terms such as AR, d and MA were also determined as 5, 1 and 1 respectively. After establishing the best fit ARIMA model, the study also needed to confirm whether the chosen model was appropriate for generating reliable forecasts. Therefore model checking phase was conducted.

3.3 Model Checking

In this phase the ADF statistical test was applied on the chosen ARIMA model to test examine behaviour of residuals around it. Based on the ADF test results it was shown that the p – value score was 4.193195e-14. This was less than the threshold critical value of 0.05. Therefore it was deduced that the series was stationary. To inspect the stability of chosen model, the residuals behaviour were further analysed in Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots. The examinations of the autocorrelation behaviour of residuals were guided by the 95% significant Upper Confidence Level (UCL) and Lower Confidence Level (LCL) bond lines. Figure 7 presents the ACF and PACF residual analysis as shown;

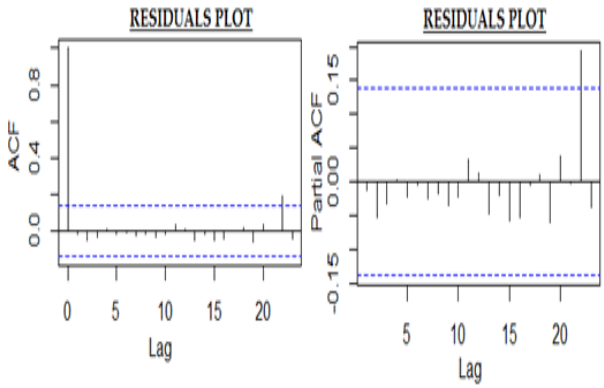

Fig. 7. The ACF and PACF Residual Analysis plots

Fig.7 shows that residuals were normally distributed with constant variance over time. However there were some significant spikes observed at lag 0 and 22 respectively. In order to confirm that the significant spikes would have no serial correlations towards the forecast intervals, a Ljung-Box statistical test was run to measure the independence and randomness of residuals on chosen best fit ARIMA model. Based on the results, it was shown that the p-value measure was 0.7814. This was more than the critical value of 0.05.Therefore this indicated that residuals were purely random with no autocorrelation with the model. Additionally, residuals also conveyed a white noise. This affirmed that the chosen model was an adequate fit to the data. Therefore it was considered for generating forecasts of COVID-19 confirmed cases in Botswana in 60 days period.

3.4 Forecasting with best ARIMA Model

In this stage the chosen ARIMA model (5,1, 1) was used to generate forecasts of COVID-19 confirmed cases forecast in 60 days period. That is from the 22nd of January 2021 to March 22nd 2021. The actual values (observed) of confirmed cases were plotted against the forecasted values. The 95% standard error bond lines (UCL and LCL) were plotted to guide the forecasts error limits. Figure 8 depicts the graphical illustration of the observed and forecast values of COVID-19 confirmed cases in Botswana over 60 days period.

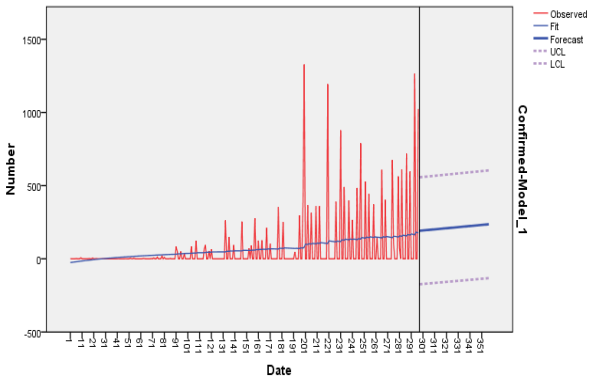

Fig. 8. Observed values vs. Forecast values of COVID-19 confirmed cases in Botswana over the next 60 days period

Fig. 8 shows that given the environmental variables remain constant, i.e. the government precaution measures, policies and other regulation used to control the virus are imposed, then confirmed cases in Botswana are expected to rise at a steady upward trend with random fluctuations and inconstant variance in the next 60 days period. In order to validate the accuracy of the forecasts, Ljung-Box Q statistical tests were performed on the forecasts residuals on different lag intervals. The statistical p-value measure was used to evaluate residuals independence and randomness. Table 1 presents the results of conducting series of Ljung-Box tests on forecasts residuals as shown;

Number of Lag runs	Ljung-Box statistical test (p- value results)
5	0.9728
10	0.9995
15	0.9997
20	0.9799
25	0.9689

Table 1 shows that residuals of the forecasts errors were normally distributed with mean close to zero. Furthermore results also showed that the residuals were significant at p-values greater than the significant interval of 0.05. These also results affirmed that that residuals were independent and conveyed no serial autocorrelations. Therefore the study failed to reject the null hypothesis. In a good forecast model, the basic assumption is that residuals should resemble zero correlations. The Ljung Box test also assumes that residuals are not significant with p value greater than zero (Fattah, Ezzine, & Aman, 2018). Based on the derived assumptions, the study concluded that the residuals of forecasts errors resembled white noise. Therefore, it was approved that the generated forecasts provided reliable estimates that could guide on understanding the future trend of confirmed cases in Botswana.

Introspecting on generated forecasts of confirmed cases for next 60 days period, much effort will be required from the government and public in order to flatten the curve of COVID-19 confirmed cases in Botswana. Therefore the study highly recommends the government should impose stricter precaution policies and measures, for instance; an execution of a second national lock down, strict curfew regulations, more strict penalties towards of compulsory wear of face mask in public areas, strict social distancing regulations and the public should also abide COVID-19 precaution measures at all the times.

4 Conclusion and Future works

The accurate forecasting of COVID-19 infectious disease had become critical for the stability of every country's economy and societal wellbeing. This study adopted an ARIMA model to forecast confirmed cases in Botswana for the next 60 days period. Findings of the pilot study suggest that ARIMA model is a powerful tool that could be used to generate forecasts which could guide on estimating future encounters with infectious diseases or pandemics. Even though there are some forecasting limitations associated with ARIMA models, the reliability and volatility of forecasts errors could be easily improved. In some cases accuracy performance metrics like a Ljung Box Q statistics test could be used.

In forecasting COVID-19 confirmed cases for the next 60 days period, pilot findings suggest that given that environmental variables remain constant (i.e. the current government precaution procedures, policies and other strategies to control the virus are imposed) then confirmed cases in any environment are expected to rise gradually in random fluctuations with a steep upward trend. Therefore in order to effectively manage the COVID-19 infections, the study recommends and support government imposition of stricter precaution policies and measures as and when necessary. This could involve execution of national lock downs, strict curfew regulations, more strict penalties towards of compulsory wearing of face mask in public areas, strict social distancing regulations and public acceptance and abiding by the COVID-19 precaution measures at all the times. This study is one of the ground-breaking studies that models forecasts COVID-19 confirmed cases in Botswana. Therefore the study findings support strategic management decisions and policy improvements to curb the COVID-19 infections in the country. It is suggested that future research should us a larger set of data covering longer periods for more reliable forecasts. Also, accuracy validation metrics such as Root Mean Squared Error, Mean Absolute Percentage Error and Mean Absolute Error could be adopted. The results could be compared to Ljung Box

statistical tests and their significance could be evaluated. Furthermore a multivariate time series algorithms such as ARIMA with exogenous regressors could be adopted. This model extends the traditional ARIMA model by taking in to account of some independent variables. Finally machine learning techniques like the artificial neural networks and forecast algorithms like Facebook prophet algorithm could be adopted to provide different ways to correlate forecasts of COVID-19 confirmed cases in Botswana.

References

[1] Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *PMC US National Library of Medicine National Institute of Health*, doi: 10.1016/j.scitotenv.2020.138817.

[2] Chae, S., Kwon, S., & Lee, D. (2018). Predicting Infectious Disease Using Deep Learning and Big Data. *International Journal of Environmental Research and Public Health*, doi: 10.3390/ijerph15081596.

[3] Chaurasia, V., & Pal, S. (2020). Application of machine learning time series analysis for prediction COVID-19 pandemic. *Springer*, https://doi.org/10.1007/s42600-020-00105-4.

[4] Fanoodi, B., Malmir, B., & Jahantigh, F. F. (2019). Reducing demand uncertainty in platelet supply chain through artificial neural networks and ARIMA models. *Elsevier Computers in Biology and Medicine* , DOI: 10.1016/j.compbiomed.2019.103415.

[5] Fattah, J., Ezzine, L., & Aman, Z. (2018). Forecasting of demand using ARIMA model . *Internationa Journal of Business Management*, https://doi.org/10.1177/1847979018808673.

[6] Hodgson, S. H., Mansatta, K., & Mallet, G. et al. (2020). What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. *The Lancert*, DOI:https://doi.org/10.1016/S1473-3099(20)30773-8.

[7] IWK Health Center. (2021, 7 January). *COVID-19 Research*. Retrieved from Library Services: https://library.nshealth.ca/COVID19Research/Publications

[8] Jia, L., Li, K., Jiang, Y., Guo, X., & zhao, T. (2020). Prediction and analysis of Coronavirus Disease 2019. *NASA Astrophysics Data System*, https://arxiv.org/ftp/arxiv/papers/2003/2003.05447.pdf.

[9] Johns Hopkins University of Medicine. (2021, January 7). *Corona virus resource center*. Retrieved from COVID-19 Dashboards by the Centerfor System Science and Engineering (CSSE) at Johns Hopkisn University: https://coronavirus.jhu.edu/map.html

[10] Koopmans, L. H. (1995). *Multivariate Spectral Models and Their Applications.* Science Direct - Elsevier.

[11] Kumar, S., & Toshniwal, D. (2016). "A data mining approach to characterize road accident locations". *J. Mod. Transport*.

[12] Our World in Data. (2020, January 7). *Statistics and Research*. Retrieved from Coronavirus Pandemic (COVID-19): https://ourworldindata.org/coronavirus#citation

[13] Steptoe, A., & Poole, L. (2015). *Infectious Diseases: Psychosocial Aspects.* International Encyclopedia of the Social & Behavioral Sciences (Second Edition).

[14] Stieg, C. (2020, March 3). *HEALTH AND WELLNESS*. Retrieved from How this Canadian start-up spotted coronavirus before everyone else knew about it: https://www.cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html

[15] Vara, V. (2020, April 16). *Pharmaceutical Technology*. Retrieved from Latest Analysis: https://www.pharmaceutical-technology.com/features/coronavirus-outbreak-the-countries-affected/

[16] Wang, M., Wang, H., Wang, J., & Lui et. al. (2019). A novel model for malaria prediction based on ensemble algorithms. *PLoS ONE*, 14(12): https://doi.org/10.1371/journal.pone.0226910.

[17] Wang, Y., Xu, C., Zhang, S., & Yang et. al. (2019). Development and evaluation of a deep learning approach formodeling seasonality and trends in hand-foot-mouth disease incidence in mainland China. *Springer Nature*.

[18] WHO. (2021, January 7). *World Health Organisation*. Retrieved from WHO Coronavirus Disease (COVID-19) Dashboard: https://covid19.who.int/

[19] World Health Organization. (2021, January 10). *WHO Health Topic Page: Zoonoses.* Retrieved from https://www.who.int/topics/zoonoses/en/

[20] Worldometer. (2021, 8 January). *COVID-19 CORONAVIRUS PANDEMIC*. Retrieved from Worldometer: https://www.worldometers.info/coronavirus/